(19)日本国特許庁(JP)　　(12)公 開 特 許 公 報(A)　　(11)特許出願公開番号
特開2020-27168
(P2020-27168A)
(43)公開日　令和2年2月20日(2020.2.20)

(51)Int.Cl.		F I		テーマコード（参考）
G10L 25/30	(2013.01)	G10L	25/30	

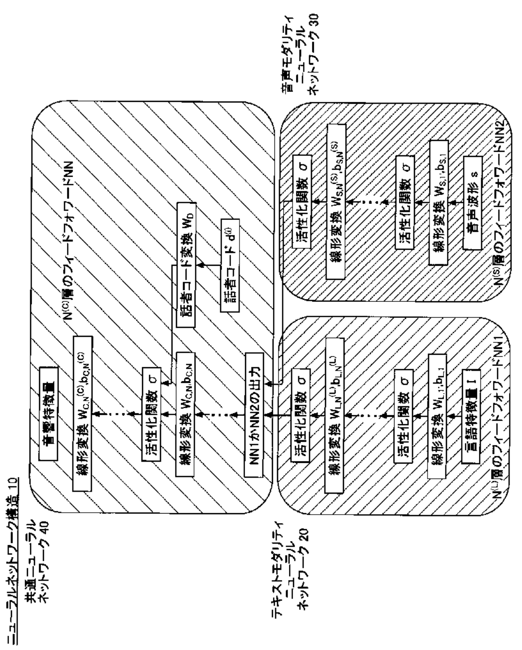
審査請求 未請求　請求項の数 11　ＯＬ　（全 19 頁）

(21)出願番号　特願2018-151611 (P2018-151611)
(22)出願日　平成30年8月10日 (2018.8.10)

(71)出願人　504202472
大学共同利用機関法人情報・システム研究機構
東京都立川市緑町１０番３号
(74)代理人　100107766
弁理士　伊東　忠重
(74)代理人　100070150
弁理士　伊東　忠彦
(72)発明者　ヒュウ　ティ　ルオン
東京都千代田区一ツ橋二丁目１番２号　大学共同利用機関法人情報・システム研究機構　国立情報学研究所内
(72)発明者　山岸　順一
東京都千代田区一ツ橋二丁目１番２号　大学共同利用機関法人情報・システム研究機構　国立情報学研究所内

(54)【発明の名称】学習装置、学習方法、音声合成装置、音声合成方法及びプログラム

(57)【要約】　（修正有）
【課題】教師有り適応と教師なし適応との何れのケースにも対応可能な、ニューラルネットワーク構造を持った未知話者のための音声合成技術を提供する。
【解決手段】テキストデータを第１のベクトルに変換するテキストモダリティニューラルネットワーク（テキストモダリティＮＮ）と、音声波形データを第２のベクトルに変換する音声モダリティＮＮと、第１又は第２のベクトルから話者空間上の話者コードベクトルに対応する音響特徴量を生成する共通ＮＮを備え、テキストデータと音響特徴量で構成される第１の訓練データによりテキストモダリティＮＮ及び共通ＮＮを学習し、音声波形データと音響特徴量で構成される第２の訓練データにより音声モダリティＮＮ及び共通ＮＮを学習し、所与の話者の第３の訓練データに応じて、テキストモダリティＮＮ及び共通ＮＮと、音声モダリティＮＮ及び共通ＮＮとを選択的に利用して、話者に対する話者コードベクトルを推定する。
【選択図】図１



【特許請求の範囲】

【請求項１】

メモリと、

プロセッサと、

を有する学習装置であって、

前記メモリは、

テキストデータを第１のベクトルに変換するテキストモダリティニューラルネットワークと、

音声波形データを第２のベクトルに変換する音声モダリティニューラルネットワークと、

前記テキストモダリティニューラルネットワーク及び前記音声モダリティニューラルネットワークに接続され、前記第１のベクトル又は前記第２のベクトルから話者空間上の話者コードベクトルに対応する音響特徴量を生成する共通ニューラルネットワークとを格納し、

前記プロセッサは、

テキストデータと音響特徴量とから構成される第１の訓練データによって前記テキストモダリティニューラルネットワーク及び前記共通ニューラルネットワークを学習し、

音声波形データと音響特徴量とから構成される第２の訓練データによって前記音声モダリティニューラルネットワーク及び前記共通ニューラルネットワークを学習し、

所与の話者の第３の訓練データに応じて、前記テキストモダリティニューラルネットワーク及び前記共通ニューラルネットワークと、前記音声モダリティニューラルネットワーク及び前記共通ニューラルネットワークとを選択的に利用して、前記所与の話者に対する前記話者コードベクトルを推定する学習装置。

【請求項２】

前記プロセッサは、

前記第１の訓練データのテキストデータを前記テキストモダリティニューラルネットワークに入力し、前記テキストモダリティニューラルネットワークから取得した第１のベクトルを前記共通ニューラルネットワークに入力し、前記共通ニューラルネットワークから取得した音響特徴量と前記第１の訓練データの音響特徴量との間の第１の誤差を算出し、

前記第２の訓練データの音声波形データを前記音声モダリティニューラルネットワークに入力し、前記音声モダリティニューラルネットワークから取得した第２のベクトルを前記共通ニューラルネットワークに入力し、前記共通ニューラルネットワークから取得した音響特徴量と前記第２の訓練データの音響特徴量との間の第２の誤差を算出し、

前記第１の誤差と前記第２の誤差との加重和に基づき、前記テキストモダリティニューラルネットワーク、前記音声モダリティニューラルネットワーク及び前記共通ニューラルネットワークを学習する、請求項１記載の学習装置。

【請求項３】

前記プロセッサは、

前記第１の訓練データのテキストデータを前記テキストモダリティニューラルネットワークに入力し、前記テキストモダリティニューラルネットワークから取得した第１のベクトルを前記共通ニューラルネットワークに入力し、前記共通ニューラルネットワークから取得した音響特徴量と前記第１の訓練データの音響特徴量との間の第１の誤差を算出し、

前記第２の訓練データの音声波形データを前記音声モダリティニューラルネットワークに入力し、前記音声モダリティニューラルネットワークから取得した第２のベクトルを前記共通ニューラルネットワークに入力し、前記共通ニューラルネットワークの一部のレイヤから構成されるサブニューラルネットワークから第３のベクトルを取得し、前記共通ニューラルネットワークに入力された第１のベクトルに対して前記サブニューラルネットワークから第４のベクトルを取得し、前記第３のベクトルと前記第４のベクトルとの間の距離に基づき第３の誤差を算出し、

前記第１の誤差と前記第３の誤差との加重和に基づき、前記テキストモダリティニュー

ラルネットワーク、前記音声モダリティニューラルネットワーク及び前記共通ニューラルネットワークを学習する、請求項１又は２記載の学習装置。

【請求項４】

前記プロセッサは、

前記第３の訓練データがテキストデータと音響特徴量とから構成される場合、前記テキストデータを前記テキストモダリティニューラルネットワークに入力し、前記テキストモダリティニューラルネットワークから取得した第１のベクトルを前記共通ニューラルネットワークに入力し、前記共通ニューラルネットワークから取得した音響特徴量と前記第３の訓練データの音響特徴量との間の第４の誤差に基づき前記所与の話者の話者コードベクトルを決定する、請求項１乃至３何れか一項記載の学習装置。

【請求項５】

前記プロセッサは、

前記第３の訓練データが音声波形データと音響特徴量とから構成される場合、前記音声波形データを前記音声モダリティニューラルネットワークに入力し、前記音声モダリティニューラルネットワークから取得した第２のベクトルを前記共通ニューラルネットワークに入力し、前記共通ニューラルネットワークから取得した音響特徴量と前記第３の訓練データの音響特徴量との間の第５の誤差に基づき前記所与の話者の話者コードベクトルを決定する、請求項１乃至４何れか一項記載の学習装置。

【請求項６】

メモリと、

プロセッサと、

を有する音声合成装置であって、

前記メモリは、

学習済みのテキストモダリティニューラルネットワークと、

所与の話者に対して学習済みの共通ニューラルネットワークと、

を格納し、

前記プロセッサは、テキストデータを取得すると、前記格納されているテキストモダリティニューラルネットワーク及び共通ニューラルネットワークによって、前記テキストデータから前記所与の話者に対応する音響特徴量を生成する音声合成装置。

【請求項７】

テキストデータを取得し、前記所与の話者に対応して前記テキストデータから生成された音響特徴量を再生する入出力インタフェースを更に有する、請求項６記載の音声合成装置。

【請求項８】

メモリとプロセッサとを有するコンピュータによって実現される学習方法であって、

前記メモリは、

テキストデータを第１のベクトルに変換するテキストモダリティニューラルネットワークと、

音声波形データを第２のベクトルに変換する音声モダリティニューラルネットワークと、

前記テキストモダリティニューラルネットワーク及び前記音声モダリティニューラルネットワークに接続され、前記第１のベクトル又は前記第２のベクトルから話者空間上の話者コードベクトルに対応する音響特徴量を生成する共通ニューラルネットワークとを格納し、

前記プロセッサが、テキストデータと音響特徴量とから構成される第１の訓練データによって前記テキストモダリティニューラルネットワーク及び前記共通ニューラルネットワークを学習するステップと、

前記プロセッサが、音声波形データと音響特徴量とから構成される第２の訓練データによって前記音声モダリティニューラルネットワーク及び前記共通ニューラルネットワークを学習するステップと、

前記プロセッサが、所与の話者の第３の訓練データに応じて、前記テキストモダリティニューラルネットワーク及び前記共通ニューラルネットワークと、前記音声モダリティニューラルネットワーク及び前記共通ニューラルネットワークとを選択的に利用して、前記所与の話者に対する前記話者コードベクトルを推定するステップと、
を有する学習方法。
【請求項９】
メモリとプロセッサとを有するコンピュータによって実現される音声合成方法であって、
前記メモリは、
学習済みのテキストモダリティニューラルネットワークと、
所与の話者に対して学習済みの共通ニューラルネットワークと、
を格納し、
前記プロセッサが、テキストデータを取得すると、前記格納されているテキストモダリティニューラルネットワーク及び共通ニューラルネットワークによって、前記テキストデータから前記所与の話者に対応する音響特徴量を生成するステップを有する音声合成方法。
【請求項１０】
テキストデータを第１のベクトルに変換するテキストモダリティニューラルネットワークと、音声波形データを第２のベクトルに変換する音声モダリティニューラルネットワークと、前記テキストモダリティニューラルネットワーク及び前記音声モダリティニューラルネットワークに接続され、前記第１のベクトル又は前記第２のベクトルから話者空間上の話者コードベクトルに対応する音響特徴量を生成する共通ニューラルネットワークとを格納したメモリに接続されるプロセッサに、
テキストデータと音響特徴量とから構成される第１の訓練データによって前記テキストモダリティニューラルネットワーク及び前記共通ニューラルネットワークを学習させ、
音声波形データと音響特徴量とから構成される第２の訓練データによって前記音声モダリティニューラルネットワーク及び前記共通ニューラルネットワークを学習させ、
所与の話者の第３の訓練データに応じて、前記テキストモダリティニューラルネットワーク及び前記共通ニューラルネットワークと、前記音声モダリティニューラルネットワーク及び前記共通ニューラルネットワークとを選択的に利用して、前記所与の話者に対する前記話者コードベクトルを推定させるプログラム。
【請求項１１】
学習済みのテキストモダリティニューラルネットワークと、所与の話者に対して学習済みの共通ニューラルネットワークとを格納したメモリに接続されるプロセッサに、
テキストデータを取得すると、前記格納されているテキストモダリティニューラルネットワーク及び共通ニューラルネットワークによって、前記テキストデータから前記所与の話者に対応する音響特徴量を生成させるプログラム。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、一般に音声合成技術に関し、より詳細には、ニューラルネットワークを利用した未知話者に対する話者適応技術に関する。
【背景技術】
【０００２】
近年のディープラーニングの進展によって、ニューラルネットワークを利用した音声合成システムの研究開発が進められている。
【０００３】
音声合成システムの一例として、特定話者のための音声合成システムがある。特定話者のための音声合成システムによると、特定話者の音声データとテキストデータとのペアを訓練データとして利用することによって、テキストデータを当該話者に対応する音声デー

タに変換するニューラルネットワークが学習され、学習済みのニューラルネットワークを利用して、入力されたテキストデータが当該特定話者の音声によって再生される。

【０００４】

他の例として、複数話者のための音声合成システムがある。複数話者のための音声合成システムによると、複数話者の音声データとテキストデータとのペアを訓練データとして利用することによって、テキストデータを複数話者の何れか指定された話者に対応する音声データに変換するニューラルネットワークが学習され、学習済みのニューラルネットワークを利用して、入力されたテキストデータが当該指定された話者の音声によって再生される。

【０００５】

更なる他の例として、未知話者のための音声合成システムがある。典型的には、上述した複数話者のための音声合成システムに基づき、未知話者の音声データ及び／又はテキストデータを訓練データとして利用することによって、テキストデータを当該未知話者に対応する音声データに変換するニューラルネットワークが学習される。学習済みのニューラルネットワークを利用して、入力されたテキストデータが当該未知話者の音声によって再生される。

【０００６】

未知話者のための音声合成システムとして、未知話者の音声データとテキストデータとのペアを訓練データとして利用するもの（教師有り適応と呼ばれる）と、未知話者の音声データのみを訓練データとして利用するもの（教師なし適応と呼ばれる）とがある。

【先行技術文献】

【非特許文献】

【０００７】

【非特許文献１】"Neural Voice Cloning with a Few Samples", Sercan O. Arik, et. al., arXiv: 1802.06006, Mar. 20, 2018.

【非特許文献２】"Fitting New Speakers Based on a Short Untranscribed Sample", Eliya Nachmani, et. al., arXiv: 1802.06984, Feb. 20, 2018.

【発明の概要】

【発明が解決しようとする課題】

【０００８】

従来技術によると、教師有り適応に基づく未知話者のための音声合成システムと、教師なし適応に基づく未知話者のための音声合成システムとは、それぞれ独立に設計されており、教師有り適応と教師なし適応との双方に対応可能な音声合成システムは現状存在しない。従って、教師有り適応と教師なし適応との何れのケースにも対応可能なニューラルネットワーク構造を備えた未知話者のための音声合成システムが望まれる。

【０００９】

上述した問題点を鑑み、本発明の課題は、教師有り適応と教師なし適応との何れのケースにも対応可能なニューラルネットワーク構造を利用した未知話者のための音声合成技術を提供することである。

【課題を解決するための手段】

【００１０】

上記課題を解決するため、本発明の一態様は、メモリと、プロセッサとを有する学習装置であって、前記メモリは、テキストデータを第１のベクトルに変換するテキストモダリティニューラルネットワークと、音声波形データを第２のベクトルに変換する音声モダリティニューラルネットワークと、前記テキストモダリティニューラルネットワーク及び前記音声モダリティニューラルネットワークに接続され、前記第１のベクトル又は前記第２のベクトルから話者空間上の話者コードベクトルに対応する音響特徴量を生成する共通ニューラルネットワークとを格納し、前記プロセッサは、テキストデータと音響特徴量とから構成される第１の訓練データによって前記テキストモダリティニューラルネットワーク及び前記共通ニューラルネットワークを学習し、音声波形データと音響特徴量とから構成

される第２の訓練データによって前記音声モダリティニューラルネットワーク及び前記共通ニューラルネットワークを学習し、所与の話者の第３の訓練データに応じて、前記テキストモダリティニューラルネットワーク及び前記共通ニューラルネットワークと、前記音声モダリティニューラルネットワーク及び前記共通ニューラルネットワークとを選択的に利用して、前記所与の話者に対する前記話者コードベクトルを推定する学習装置に関する。

【発明の効果】

【００１１】

本発明によると、教師有り適応と教師なし適応との何れのケースにも対応可能なニューラルネットワーク構造を利用した未知話者のための音声合成技術を提供することができる。

【図面の簡単な説明】

【００１２】

【図１】本発明の一実施例によるニューラルネットワーク構造の概略図である。

【図２】本発明の一実施例による学習装置及び音声合成装置のハードウェア構成を示すブロック図である。

【図３】本発明の一実施例による学習処理を示す概略図である。

【図４】本発明の一実施例による学習処理を示すフローチャートである。

【図５】本発明の他の実施例による学習処理を示す概略図である。

【図６】本発明の他の実施例による学習処理を示すフローチャートである。

【図７】本発明の一実施例による未知話者適応処理を示す概略図である。

【図８】本発明の一実施例による未知話者適応処理を示すフローチャートである。

【図９】本発明の一実施例による音声合成処理を示す概略図である。

【図１０】本発明の一実施例による音声合成処理を示すフローチャートである。

【図１１】本発明の各種実施例による学習処理の実験結果を示す図である。

【発明を実施するための形態】

【００１３】

以下の実施例では、教師有り適応と教師なし適応との何れのケースにも対応可能なニューラルネットワークを学習する学習装置１００と、当該ニューラルネットワークを利用した未知話者のための音声合成装置２００とが開示される。

［概略］

後述される実施例を概略すると、学習装置１００は、テキストデータをベクトルに変換するテキストモダリティニューラルネットワーク２０、音声波形データをベクトルに変換する音声モダリティニューラルネットワーク３０、及びテキストモダリティニューラルネットワーク２０及び音声モダリティニューラルネットワーク３０から出力されたベクトルから、話者空間上の所与の未知話者を示す話者コードベクトル（潜在変数）に対応する音響特徴量を生成する共通ニューラルネットワーク４０を学習する。

【００１４】

まず、テキストモダリティニューラルネットワーク２０、音声モダリティニューラルネットワーク３０及び共通ニューラルネットワーク４０から構成されるニューラルネットワーク構造１０に対する学習処理において、学習装置１００は、テキストデータと音響特徴量とのペアから構成される訓練データに対して、テキストデータをテキストモダリティニューラルネットワーク２０に入力し、テキストモダリティニューラルネットワーク２０から出力されたベクトルを共通ニューラルネットワーク４０に入力する。一方、学習装置１００は、音声波形データと音響特徴量とのペアから構成される訓練データに対して、音声波形データを音声モダリティニューラルネットワーク３０に入力し、音声モダリティニューラルネットワーク３０から取得されたベクトルを共通ニューラルネットワーク４０に入力する。そして、以下の実施例において詳細に説明されるように、学習装置１００は、共通ニューラルネットワーク４０から出力された音響特徴量と訓練データの音響特徴量とに基づき、テキストモダリティニューラルネットワーク２０、音声モダリティニューラルネットワーク

ットワーク３０及び共通ニューラルネットワーク４０を学習する。

【００１５】

次に、未知話者適応処理において、学習装置１００は、上述した学習済みのテキストモダリティニューラルネットワーク２０、音声モダリティニューラルネットワーク３０及び共通ニューラルネットワーク４０を利用して、話者空間上の未知話者の位置を示す話者コードベクトルを推定する。すなわち、所与の話者の訓練データが与えられると、学習装置１００は、当該訓練データがテキスト付きの音声データであるか、あるいは、音声データのみであるかに応じて、テキストモダリティニューラルネットワーク２０及び共通ニューラルネットワーク４０と、音声モダリティニューラルネットワーク３０及び共通ニューラルネットワーク４０とを選択的に利用して、共通ニューラルネットワーク４０の話者空間上の当該話者を示す話者コードベクトル（潜在変数）を推定し、推定した潜在変数が埋め込まれた話者毎の共通ニューラルネットワーク４０を生成する。

【００１６】

音声合成装置２００は、このようにして学習装置１００によって未知話者毎に学習されたニューラルネットワーク構造１０における学習済みのテキストモダリティニューラルネットワーク２０及び共通ニューラルネットワーク４０を利用して、所与のテキストデータから当該未知話者に対応する音声データを生成する。

［ニューラルネットワーク構造］

まず、図１を参照して、本発明の一実施例によるニューラルネットワーク構造１０を説明する。図１は、本発明の一実施例によるニューラルネットワーク構造１０の概略図である。

【００１７】

図１に示されるように、ニューラルネットワーク構造１０は、テキストモダリティニューラルネットワーク２０、音声モダリティニューラルネットワーク３０及び共通ニューラルネットワーク４０を有し、テキストモダリティニューラルネットワーク２０及び音声モダリティニューラルネットワーク３０はそれぞれ、共通ニューラルネットワーク４０に接続される。

【００１８】

テキストモダリティニューラルネットワーク２０は、入力されたテキストデータ（例えば、言語特徴量）を共通ニューラルネットワーク４０への入力用のベクトルに変換する何れかのレイヤ構成を有するニューラルネットワークである。図示された実施例では、テキストモダリティニューラルネットワーク２０は、$N^{(L)}$層のフィードフォワードニューラルネットワークであり、テキストデータのベクトルlを入力層において取得し、取得したベクトルlを隠れ層にわたす。$N^{(L)}$個の隠れ層はそれぞれ、前段のレイヤからわたされたベクトルを行列W_L及びバイアスベクトルb_Lによって線形変換し、変換されたベクトルを活性化関数σ（例えば、シグモイド関数）に入力し、活性化関数σから出力されたベクトルを後段のレイヤにわたす。出力層は、前段の隠れ層からわたされたベクトルを共通ニューラルネットワーク４０の入力層にわたす。

【００１９】

形式的には、テキストデータのベクトルlが与えられると、第１の隠れ層は、

$$h_1 = \sigma(W_{L,1} l + b_{L,1})$$

によってベクトルh_1を出力する。以下同様にして、各隠れ層は同様の変換処理を実行し、第$N^{(L)}$の隠れ層は、前段の隠れ層からベクトルh_{NL-1}が与えられると、

$$h_{NL} = \sigma(W_{L,NL} h_{NL-1} + b_{L,NL})$$

によってベクトルh_{NL}を出力し、出力層にわたす。当該ベクトル及び行列は、後述される学習処理において学習される。

【００２０】

音声モダリティニューラルネットワーク３０は、入力された音声データ（例えば、音声波形）を共通ニューラルネットワーク４０への入力用のベクトルに変換する何れかのレイヤ構成を有するニューラルネットワークである。図示された実施例では、音声モダリティ

ニューラルネットワーク３０は、$N^{(S)}$層のフィードフォワードニューラルネットワークであり、音声データのベクトルｓを入力層において取得し、取得したベクトルｓを隠れ層にわたす。$N^{(S)}$個の隠れ層はそれぞれ、前段のレイヤからわたされたベクトルを行列W_S及びバイアスベクトルb_Sによって線形変換し、変換されたベクトルを活性化関数σ（例えば、シグモイド関数）に入力し、活性化関数σから出力されたベクトルを後段のレイヤにわたす。出力層は、前段の隠れ層からわたされたベクトルを共通ニューラルネットワーク４０の入力層にわたす。なお、各隠れ層における具体的な処理は、上述したテキストモダリティニューラルネットワーク２０のものと同様であり、重複する説明は省く。

【００２１】

共通ニューラルネットワーク４０は、テキストモダリティニューラルネットワーク２０及び音声モダリティニューラルネットワーク３０からわたされたベクトルを音響特徴量に変換する何れかのレイヤ構成を有するニューラルネットワークである。図示された実施例では、共通ニューラルネットワーク４０は、$N^{(C)}$層のフィードフォワードニューラルネットワークであり、テキストモダリティニューラルネットワーク２０及び音声モダリティニューラルネットワーク３０から入力されたベクトルを入力層において取得し、取得したベクトルを隠れ層にわたす。$N^{(C)}$個の隠れ層はそれぞれ、前段のレイヤからわたされたベクトルを行列W_C及びバイアスベクトルb_Cによって線形変換し、変換されたベクトルを活性化関数σ（例えば、シグモイド関数）に入力し、活性化関数σから出力されたベクトルを後段のレイヤにわたす。出力層は、前段の隠れ層からわたされた音響特徴量を示すベクトルを出力する。

【００２２】

また、共通ニューラルネットワーク４０は更に、後述される未知話者適応処理によって推定された所与の話者を示す話者コードベクトル（潜在変数）を含む。換言すると、共通ニューラルネットワーク４０は、未知話者適応処理において学習装置１００によって話者毎に学習される。所与の話者を示す話者空間上の推定された話者コードベクトルが与えられた隠れ層は、前段のレイヤからわたされたベクトルと話者コードベクトルとに対して線形変換を実行し、変換されたベクトルを活性化関数σ（例えば、シグモイド関数）に入力し、活性化関数σから出力されたベクトルを後段のレイヤにわたす。

【００２３】

形式的には、話者コードベクトルが与えられる隠れ層は、前段のレイヤからベクトルh_{n-1}と話者ｉの話者コードベクトル$d^{(i)}$とが与えられると、

$$h_n = \sigma(W_{C,n} h_{n-1} + b_{C,n} + W_D d^{(i)})$$

によってベクトルh_nを取得する。ここで、W_Dは話者コード用の重み行列である。なお、話者コードベクトルが入力されない各隠れ層における具体的な処理は、上述したテキストモダリティニューラルネットワーク２０のものと同様であり、重複する説明は省く。

【００２４】

なお、図示された実施例では、話者コードベクトルは１つの隠れ層にわたされているが、これに限定されるものでなく、共通ニューラルネットワーク４０のレイヤ構成に応じて複数の隠れ層にわたされてもよい。

［ハードウェア構成］

ここで、学習装置１００及び音声合成装置２００は、例えば、図２に示されるように、CPU (Central Processing unit)、GPU (Graphics Processing Unit)などのプロセッサ１０１、RAM (Random Access Memory)、フラッシュメモリなどのメモリ１０２、ハードディスク１０３及び入出力(I/O)インタフェース１０４によるハードウェア構成を有してもよい。

【００２５】

プロセッサ１０１は、学習装置１００及び音声合成装置２００の各種処理を実行する。

【００２６】

メモリ１０２は、学習装置１００及び音声合成装置２００における各種データ及びプログラムを格納し、特に作業用データ、実行中のプログラムなどのためのワーキングメモリ

として機能する。具体的には、メモリ１０２は、ハードディスク１０３からロードされたニューラルネットワーク構造１０を実現するプログラム、各種処理を実行及び制御するためのプログラムなどを格納し、プロセッサ１０１によるプログラムの実行中にワーキングメモリとして機能する。

【００２７】

ハードディスク１０３は、学習装置１００及び音声合成装置２００における各種データ及びプログラムを格納する。

【００２８】

I/Oインタフェース１０４は、ユーザからの命令、入力データなどを受け付け、出力結果を表示、再生などすると共に、外部装置との間でデータを入出力するためのインタフェースである。例えば、I/Oインタフェース１０４は、USB (Universal Serial Bus)、通信回線、キーボード、マウス、ディスプレイ、マイクロフォン、スピーカなどの各種データを入出力するためのデバイスである。

【００２９】

しかしながら、本発明による学習装置１００及び音声合成装置２００は、上述したハードウェア構成に限定されず、他の何れか適切なハードウェア構成を有してもよい。例えば、上述した学習装置１００及び音声合成装置２００による各種処理の１つ以上は、これを実現するよう配線化された処理回路又は電子回路により実現されてもよい。

［ニューラルネットワーク構造の第１の学習処理］

次に、図３及び４を参照して、本発明の一実施例によるニューラルネットワーク構造１０に対する学習処理を説明する。上述したニューラルネットワーク構造１０の内部構成から理解されるように、学習装置１００は、共通ニューラルネットワーク４０がテキストデータと音声データとの異なるモダリティからの入力を適切に受け付けるようにニューラルネットワーク構造１０を学習する必要がある。

【００３０】

図３は、本発明の一実施例による学習処理を示す概略図である。本実施例では、図３に示されるように、学習装置１００は、共通ニューラルネットワーク４０をテキストモダリティニューラルネットワーク２０と音声モダリティニューラルネットワーク３０とに共有させ、２つの共通ニューラルネットワーク４０を同時に、すなわち、２つの共通ニューラルネットワーク４０におけるパラメータ（例えば、隠れ層の重み行列）が同一となるよう同期的に学習する。

【００３１】

具体的には、学習装置１００は、テキストデータと音響特徴量とのペアから構成される訓練データに対して、当該テキストデータをテキストモダリティニューラルネットワーク２０に入力し、テキストモダリティニューラルネットワーク２０から出力されたベクトルを取得する。そして、学習装置１００は、取得したベクトルを共通ニューラルネットワーク４０に入力し、共通ニューラルネットワーク４０から出力された音響特徴量を取得し、取得した音響特徴量と訓練データの音響特徴量との間の誤差（loss_{main}）を算出する。

【００３２】

一方、学習装置１００は、音声波形データと音響特徴量とのペアから構成される訓練データに対して、当該音声波形データを音声モダリティニューラルネットワーク３０に入力し、音声モダリティニューラルネットワーク３０から出力されたベクトルを取得する。そして、学習装置１００は、取得したベクトルを共通ニューラルネットワーク４０に入力し、共通ニューラルネットワーク４０から出力された音響特徴量を取得し、取得した音響特徴量と訓練データの音響特徴量との間の誤差（loss_{sub}）を算出する。

【００３３】

その後、学習装置１００は、算出した２つの誤差（loss_{main}，loss_{sub}）の加重和に基づき、テキストモダリティニューラルネットワーク２０、音声モダリティニューラルネットワーク３０及び共通ニューラルネットワーク４０を学習する。例えば、

学習装置１００は、
$$\mathrm{loss} = \mathrm{loss}_{main} + \alpha\, \mathrm{loss}_{sub}$$
に従って（αは、スカラー値である）、テキストモダリティニューラルネットワーク２０及び共通ニューラルネットワーク４０による誤差loss_{main}と、音声モダリティニューラルネットワーク３０及び共通ニューラルネットワーク４０による誤差loss_{sub}との２つの誤差の加重和（ｌｏｓｓ）を算出してもよい。
【００３４】
学習装置１００は、算出した誤差の加重和（ｌｏｓｓ）が減少するように、例えば、バックプロパゲーションに従って、共有される２つの共通ニューラルネットワーク４０のパラメータが同一となるように、テキストモダリティニューラルネットワーク２０、音声モダリティニューラルネットワーク３０及び共通ニューラルネットワーク４０のパラメータ（例えば、隠れ層の重み行列）を更新する。
【００３５】
図４は、本発明の一実施例による学習処理を示すフローチャートである。当該学習処理は、学習装置１００、具体的には、学習装置１００のプロセッサ１０１によって実行される。
【００３６】
図４に示されるように、ステップＳ１０１において、学習装置１００は、訓練データを取得する。例えば、訓練データが複数の話者によるテキスト付きの音声データである場合、学習装置１００は、前処理として、当該音声データを対応する音声波形データ及び音響特徴量に変換し、訓練データからテキストデータと音響特徴量とのペアと音声波形データと音響特徴量とのペアとを生成してもよい。
【００３７】
ステップＳ１０２において、学習装置１００は、処理対象の訓練データがテキストデータと音響特徴量とのペアである場合、ステップＳ１０３に進み、処理対象の訓練データが音声波形データと音響特徴量とのペアである場合、ステップＳ１０６に進む。
【００３８】
ステップＳ１０３において、学習装置１００は、訓練データのテキストデータをテキストモダリティニューラルネットワーク２０に入力し、テキストモダリティニューラルネットワーク２０から出力されたベクトルを取得する。
【００３９】
ステップＳ１０４において、学習装置１００は、取得したベクトルを共通ニューラルネットワーク４０に入力し、共通ニューラルネットワーク４０から出力された音響特徴量を取得する。
【００４０】
ステップＳ１０５において、学習装置１００は、共通ニューラルネットワーク４０から取得した音響特徴量と訓練データの音響特徴量との誤差（loss_{main}）を算出する。
【００４１】
一方、ステップＳ１０６において、学習装置１００は、訓練データの音声波形データを音声モダリティニューラルネットワーク３０に入力し、音声モダリティニューラルネットワーク３０から出力されたベクトルを取得する。
【００４２】
ステップＳ１０７において、学習装置１００は、取得したベクトルを共通ニューラルネットワーク４０に入力し、共通ニューラルネットワーク４０から出力された音響特徴量を取得する。
【００４３】
ステップＳ１０８において、学習装置１００は、共通ニューラルネットワーク４０から取得した音響特徴量と訓練データの音響特徴量との誤差（loss_{sub}）を算出する。
【００４４】

ステップＳ１０９において、学習装置１００は、ステップＳ１０５及びＳ１０８において取得した２つの誤差の加重和（ｌｏｓｓ）を計算し、計算した加重和（ｌｏｓｓ）が減少するように、例えば、バックプロパゲーションに従ってテキストモダリティニューラルネットワーク２０、音声モダリティニューラルネットワーク３０及び共通ニューラルネットワーク４０のパラメータ（例えば、隠れ層の重み行列）を更新し、具体的には、共有される２つの共通ニューラルネットワーク４０のパラメータが同一のものに更新されるように、２つの共通ニューラルネットワーク４０を同期的に学習する。

【００４５】

学習装置１００は、所定の終了条件を充足するまで、各訓練データに対して上述したステップＳ１０１～Ｓ１０９を繰り返す。当該所定の終了条件は、例えば、所定の回数の繰り返しを終了したこと、誤差（ｌｏｓｓ）が所定の閾値以下になったこと、誤差（ｌｏｓｓ）が収束したことなどであってもよい。

［ニューラルネットワーク構造の第２の学習処理］

次に、図５及び６を参照して、本発明の他の実施例によるニューラルネットワーク構造１０に対する学習処理を説明する。上述したニューラルネットワーク構造１０から理解されるように、学習装置１００は、共通ニューラルネットワーク４０がテキストデータと音声データとの異なるモダリティからの入力を適切に受け付けるようにニューラルネットワーク構造１０、特に、共通ニューラルネットワーク４０の入力層に近い下層レイヤを学習することが求められる。

【００４６】

図５は、本発明の他の実施例による学習処理を示す概略図である。本実施例では、図５に示されるように、学習装置１００は、テキストモダリティニューラルネットワーク２０及び音声モダリティニューラルネットワーク３０から入力された各ベクトルに対して、共通ニューラルネットワーク４０における一部の隠れ層（例えば、入力層から所定番目の隠れ層）から出力される各ベクトルの間の距離を損失又はペナルティ（loss_{sub}）として利用し、上述したテキストモダリティニューラルネットワーク２０及び共通ニューラルネットワーク４０における誤差（loss_{main}）と、一部の隠れ層から出力されるベクトル間の距離（loss_{sub}）との加重和に基づき、テキストモダリティニューラルネットワーク２０、音声モダリティニューラルネットワーク３０及び共通ニューラルネットワーク４０を学習する。図３及び４を参照して上述した実施例による学習処理では、隠れ層の重み行列は共有される共通ニューラルネットワーク４０において同じとされたが、テキストモダリティニューラルネットワーク２０及び音声モダリティニューラルネットワーク３０から入力された各ベクトルに対する共通ニューラルネットワーク４０の隠れ層から出力されるベクトルが互いに近いものになることを明示的に保証するものでない。このため、共通ニューラルネットワーク４０の入力層に隠れ層から出力されるベクトルが近似したものになるよう共通ニューラルネットワーク４０を学習することによって、より精度の高い変換が可能になると考えられる。

【００４７】

具体的には、学習装置１００は、テキストデータと音響特徴量とのペアから構成される訓練データに対して、当該テキストデータをテキストモダリティニューラルネットワーク２０に入力し、テキストモダリティニューラルネットワーク２０から出力されたベクトルを取得する。そして、学習装置１００は、取得したベクトルを共通ニューラルネットワーク４０に入力し、共通ニューラルネットワーク４０から出力された音響特徴量を取得し、取得した音響特徴量と訓練データの音響特徴量との間の誤差（loss_{main}）を算出する。

【００４８】

一方、学習装置１００は、音声波形データと音響特徴量とのペアから構成される訓練データに対して、当該音声波形データを音声モダリティニューラルネットワーク３０に入力し、音声モダリティニューラルネットワーク３０から出力されたベクトルを取得する。そして、学習装置１００は、取得したベクトルを共通ニューラルネットワーク４０に入力し

、共通ニューラルネットワーク４０の一部のレイヤ（例えば、入力層からＬ番目の隠れ層）から構成されるサブニューラルネットワークから出力されたベクトル（h^{l}_{sub}）を取得する一方、テキストモダリティニューラルネットワーク２０から共通ニューラルネットワーク４０に入力されたベクトルに対して、当該サブニューラルネットワークから出力されたベクトル（h^{l}_{main}）を取得する。

【００４９】

その後、学習装置１００は、２つのベクトル（h^{l}_{main}，h^{l}_{sub}）の間の距離に基づき誤差（$loss_{sub}$）を算出し、誤差（$loss_{main}$）と誤差（$loss_{sub}$）との加重和に基づきテキストモダリティニューラルネットワーク２０、音声モダリティニューラルネットワーク３０及び共通ニューラルネットワーク４０を学習する。例えば、学習装置１００は、

$$loss = loss_{main} + \beta \Sigma_{l}^{L} distance(h^{l}_{main}, h^{l}_{sub})$$

に従って（βは、スカラー値である）、２つの誤差（$loss_{main}$，$loss_{sub}$）の加重和ｌｏｓｓを算出してもよい。ここで、距離ｄｉｓｔａｎｃｅは、例えば、コサイン距離であってもよい。

【００５０】

学習装置１００は、算出した誤差の加重和が減少するように、例えば、バックプロパゲーションに従ってテキストモダリティニューラルネットワーク２０、音声モダリティニューラルネットワーク３０及び共通ニューラルネットワーク４０のパラメータ（例えば、隠れ層の重み行列）を更新する。

【００５１】

図６は、本発明の他の実施例による学習処理を示すフローチャートである。当該学習処理は、学習装置１００、具体的には、学習装置１００のプロセッサ１０１によって実行される。

【００５２】

図６に示されるように、ステップＳ２０１において、学習装置１００は、訓練データを取得する。

【００５３】

ステップＳ２０２において、学習装置１００は、処理対象の訓練データがテキストデータと音響特徴量とのペアである場合、ステップＳ２０３に進み、処理対象の訓練データが音声波形データと音響特徴量とのペアである場合、ステップＳ２０６に進む。

【００５４】

ステップＳ２０３において、学習装置１００は、訓練データのテキストデータをテキストモダリティニューラルネットワーク２０に入力し、テキストモダリティニューラルネットワーク２０から出力されたベクトルを取得する。

【００５５】

ステップＳ２０４において、学習装置１００は、取得したベクトルを共通ニューラルネットワーク４０に入力する。

【００５６】

ステップＳ２０５において、学習装置１００は、共通ニューラルネットワーク４０から出力された音響特徴量を取得すると共に、共通ニューラルネットワーク４０のサブニューラルネットワーク（例えば、入力層から所定番目の隠れ層）から出力されたベクトル（h^{l}_{main}）を取得する。

【００５７】

一方、ステップＳ２０６において、学習装置１００は、訓練データの音声波形データを音声モダリティニューラルネットワーク３０に入力し、音声モダリティニューラルネットワーク３０から出力されたベクトルを取得する。

【００５８】

ステップＳ２０７において、学習装置１００は、取得したベクトルを共通ニューラルネットワーク４０に入力する。

ットワーク４０に入力する。

【００５９】

ステップＳ２０８において、学習装置１００は、共通ニューラルネットワーク４０のサブニューラルネットワークから出力されたベクトル（h^{l}_{sub}）を取得する。

【００６０】

ステップＳ２０９において、学習装置１００は、共通ニューラルネットワーク４０から取得した音響特徴量と訓練データの音響特徴量との誤差（$loss_{main}$）と、２つのベクトル（h^{l}_{main}，h^{l}_{sub}）の間の距離（$loss_{sub}$）とを算出する。

【００６１】

ステップＳ２１０において、学習装置１００は、ステップＳ２０９において算出した誤差（$loss_{main}$）と距離（$loss_{sub}$）との加重和（ｌｏｓｓ）を算出し、算出した加重和（ｌｏｓｓ）が減少するように、例えば、バックプロパゲーションに従ってテキストモダリティニューラルネットワーク２０、音声モダリティニューラルネットワーク３０及び共通ニューラルネットワーク４０のパラメータ（例えば、隠れ層の重み行列）を更新する。

【００６２】

学習装置１００は、所定の終了条件を充足するまで、各訓練データに対して上述したステップＳ２０１～Ｓ２１０を繰り返す。当該所定の終了条件は、例えば、所定の回数の繰り返しを終了したこと、誤差（ｌｏｓｓ）が所定の閾値以下になったこと、誤差（ｌｏｓｓ）が収束したことなどであってもよい。

【００６３】

なお、２つのタイプのニューラルネットワーク構造の学習処理について個別に説明したが、これら２つのタイプの学習処理が組み合わせ可能であることは当業者に理解されるであろう。この場合、誤差（ｌｏｓｓ）は、例えば、

$$loss = loss_{main} + \alpha\, loss_{sub} + \beta \Sigma_{l}^{L} distance(h^{l}_{main}, h^{l}_{sub})$$

に従って算出され、テキストモダリティニューラルネットワーク２０、音声モダリティニューラルネットワーク３０及び共通ニューラルネットワーク４０のパラメータが、誤差を減少させるように更新されると共に、２つの共通ニューラルネットワーク４０のパラメータが同期的に学習される。

［共通ニューラルネットワーク４０に対する話者適応処理］

次に、図７及び８を参照して、本発明の一実施例による共通ニューラルネットワーク４０に対する話者適応処理を説明する。本実施例では、上述した学習処理に従ってニューラルネットワーク構造１０を学習した後、所与の未知話者の訓練データが与えられると、学習装置１００は、当該訓練データに応じて、テキストモダリティニューラルネットワーク２０及び共通ニューラルネットワーク４０と、音声モダリティニューラルネットワーク及び共通ニューラルネットワーク４０とを選択的に利用して、共通ニューラルネットワーク４０の話者空間における当該未知話者を示す話者コードベクトルを推定する。

【００６４】

図７は、本発明の一実施例による未知話者適応処理を示す概略図である。本実施例では、図７に示されるように、与えられた訓練データが所与の未知話者のテキストデータと音響特徴量とのペアである場合、学習装置１００は、テキストモダリティニューラルネットワーク２０及び共通ニューラルネットワーク４０を利用して、当該未知話者の話者コードベクトルを推定する。他方、与えられた訓練データが所与の未知話者の音声波形データと音響特徴量とのペアである場合、学習装置１００は、音声モダリティニューラルネットワーク３０及び共通ニューラルネットワーク４０を利用して、当該未知話者の話者コードベクトルを推定する。

【００６５】

具体的には、学習装置１００は、所与の未知話者の訓練データがテキストデータと音響特徴量とから構成される場合、当該テキストデータをテキストモダリティニューラルネットワ

トワーク２０に入力し、テキストモダリティニューラルネットワーク２０から取得したベクトルを共通ニューラルネットワーク４０に入力し、共通ニューラルネットワーク４０から取得した音響特徴量と訓練データの音響特徴量との間の誤差に基づき当該話者の話者コードベクトルを決定する。他方、学習装置１００は、所与の未知話者の訓練データが音声波形データと音響特徴量とから構成される場合、音声波形データを音声モダリティニューラルネットワーク３０に入力し、音声モダリティニューラルネットワーク３０から取得したベクトルを共通ニューラルネットワーク４０に入力し、共通ニューラルネットワーク４０から取得した音響特徴量と訓練データの音響特徴量との間の誤差に基づき当該話者の話者コードベクトルを決定する。

【００６６】

例えば、図１に示される具体例によると、話者コードベクトル$d^{(i)}$は、

$$d^{(i)} = d^{(i)} + \varepsilon W_D^{T} f_{n-1}$$

に従って更新される。ここで、εは所定値以下の小さな値であり、ｆは誤差伝搬のための関数であり、

$$f_{N-1}^{(C)} = W_{C,N}^{(C),T} \sigma^{-1}(e')$$

として定義され、σ^{-1}は活性化関数によって決定される伝搬用の関数であり、ｅ'は共通ニューラルネットワーク４０から取得した音響特徴量と訓練データの音響特徴量との間の誤差の微分値である。なお、当該未知話者適応処理では、共通ニューラルネットワーク４０の重み行列Ｗ及びバイアスベクトルｂは更新されない。

【００６７】

このようにして、共通ニューラルネットワーク４０における話者コードベクトル（潜在変数）を特定することによって、学習済みのニューラルネットワーク構造１０を特定の未知話者に適応させることができる。

【００６８】

図８は、本発明の一実施例による未知話者適応処理を示すフローチャートである。当該学習処理は、学習装置１００、具体的には、学習装置１００のプロセッサ１０１によって実行される。

【００６９】

図８に示されるように、ステップＳ３０１において、学習装置１００は、所与の未知話者の訓練データを取得する。

【００７０】

ステップＳ３０２において、学習装置１００は、訓練データがテキストデータと音響特徴量とのペア又は音声波形データと音響特徴量とのペアから構成されているか判断し、訓練データがテキストデータと音響特徴量とのペアから構成されている場合、ステップＳ３０３に進み、訓練データが音声波形データと音響特徴量とのペアから構成されている場合、ステップＳ３０６に進む。

【００７１】

ステップＳ３０３において、学習装置１００は、訓練データのテキストデータをテキストモダリティニューラルネットワーク２０に入力し、テキストモダリティニューラルネットワーク２０から出力されたベクトルを取得する。

【００７２】

ステップＳ３０４において、学習装置１００は、取得したベクトルを共通ニューラルネットワーク４０に入力し、共通ニューラルネットワーク４０から出力された音響特徴量を取得する。

【００７３】

ステップＳ３０５において、学習装置１００は、共通ニューラルネットワーク４０から取得した音響特徴量と訓練データの音響特徴量との間の誤差を算出する。

【００７４】

一方、ステップＳ３０６において、学習装置１００は、訓練データの音声波形データを音声モダリティニューラルネットワーク３０に入力し、音声モダリティニューラルネットワーク３０に入力し、音声モダリティニューラルネット

ワーク３０から出力されたベクトルを取得する。

【００７５】

ステップＳ３０７において、学習装置１００は、取得したベクトルを共通ニューラルネットワーク４０に入力し、共通ニューラルネットワーク４０から出力された音響特徴量を取得する。

【００７６】

ステップＳ３０８において、学習装置１００は、共通ニューラルネットワーク４０から取得した音響特徴量と訓練データの音響特徴量との間の誤差を算出する。

【００７７】

ステップＳ３０９において、学習装置１００は、ステップＳ３０５及びＳ３０８において算出した誤差が減少するように、例えば、上述した更新式を利用してバックプロパゲーションに従って共通ニューラルネットワーク４０の話者コードベクトルを更新する。

【００７８】

学習装置１００は、所定の終了条件を充足するまで、各訓練データに対して上述したステップＳ３０１～Ｓ３０９を繰り返す。当該所定の終了条件は、例えば、所定の回数の繰り返しを終了したこと、誤差が所定の閾値以下になったこと、誤差が収束したことなどであってもよい。

［学習済みニューラルネットワーク構造を利用した音声合成処理］

次に、図９～１１を参照して、本発明の一実施例による音声合成処理を説明する。本実施例では、音声合成装置２００は、上述した学習装置１００によって特定の話者に対して学習されたテキストモダリティニューラルネットワーク２０及び共通ニューラルネットワーク４０を利用して、音声合成対象のテキストデータから当該話者に対応する音声データを生成及び再生する。

【００７９】

図９は、本発明の一実施例による音声合成処理を示す概略図である。本実施例では、音声合成装置２００は、音声合成対象のテキストデータが与えられると、図９に示されるように、上述した学習装置１００によって特定の話者に対して学習されたテキストモダリティニューラルネットワーク２０及び共通ニューラルネットワーク４０を利用して、当該テキストデータから当該話者に対応する音響特徴量を生成する。具体的には、音声合成装置２００は、入出力インタフェース１０４を介して、テキストデータを取得し、当該話者に対応するテキストデータから生成された音響特徴量を再生してもよい。

【００８０】

図１０は、本発明の一実施例による音声合成処理を示すフローチャートである。当該音声合成処理は、音声合成装置２００、具体的には、音声合成装置２００のプロセッサ１０１によって実行される。

【００８１】

図１０に示されるように、ステップＳ４０１において、音声合成装置２００は、音声合成対象となるテキストデータを取得する。例えば、テキストデータは、音声合成装置２００の入出力インタフェース１０４を介し入力されたものであってもよい。

【００８２】

ステップＳ４０２において、音声合成装置２００は、取得したテキストデータを学習済みテキストモダリティニューラルネットワーク２０に入力し、テキストモダリティニューラルネットワーク２０から出力されたベクトルを取得する。

【００８３】

ステップＳ４０３において、音声合成装置２００は、取得したベクトルを学習済み共通ニューラルネットワーク４０に入力し、共通ニューラルネットワーク４０から出力された音響特徴量を取得する。

【００８４】

ステップＳ４０４において、音声合成装置２００は、共通ニューラルネットワーク４０から取得した特定の話者に対応する音響特徴量を何れかの音声データフォーマットに変換

し、変換された音声データを再生する。例えば、変換された音声データは、当該話者の声、テンポ、アクセントなどに近い音声によって入力されたテキストデータを再生したものとなりうる。

【００８５】

図１１は、本発明の各種実施例による学習処理の実験結果を示す図である。図１１において、ＶＬ、ＳＳ、ＪＧ、ＴＬ及びＪＧ＋ＴＬは、上述した学習済みニューラルネットワーク構造を利用したものを含む各種音声合成システムを表す。

【００８６】

ＶＬは、３つのニューラルネットワークから構成されるニューラルネットワーク構造１０でなく、従来のニューラルネットワーク構造を利用したシステムである。ＳＳは、ニューラルネットワーク構造１０の各モダリティニューラルネットワークを単純に置き換えて学習されたシステムである。ＪＧは、図３及び４を参照して説明した学習処理により学習されたニューラルネットワーク構造１０を利用したシステムである。ＴＬは、図５及び６を参照して説明した学習処理により学習されたニューラルネットワーク構造１０を利用したシステムである。ＪＧ＋ＴＬは、ＪＧとＴＬとを組み合わせた学習処理により学習されたニューラルネットワーク構造１０を利用したシステムである。

【００８７】

図１１では、１０、４０、１６０及び３２０個の未知話者の訓練データによって学習された場合の各音声合成システムの誤差（ＭＣＤ）のシミュレーション結果が示される。図から理解されうるように、訓練データとして音声データとテキストデータとが与えられる教師有り学習と、音声データのみが与えられる教師なし学習との何れのケースでも、上述した実施例によるＪＧ、ＴＬ及びＪＧ＋ＴＬは、ＶＬ及びＳＳに対して有意に誤差を低減するという結果を得ることができた。

【００８８】

なお、上述した実施例では、テキストデータと音声データとが異なるモダリティとして扱われたが、本発明は、これに限定されるものでなく、他のタイプのモダリティの組み合わせに同様にして適用可能であることは理解されるであろう。

【００８９】

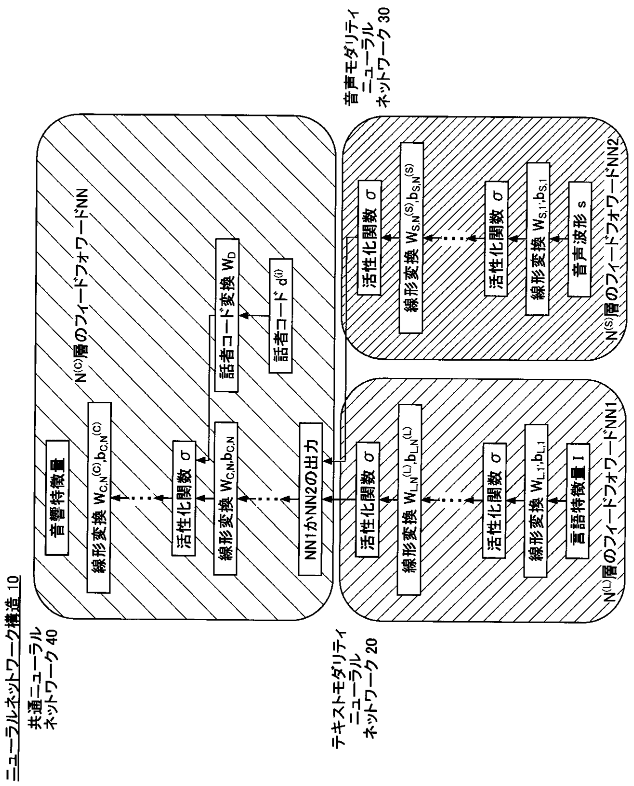
以上、本発明の実施例について詳述したが、本発明は上述した特定の実施形態に限定されるものではなく、特許請求の範囲に記載された本発明の要旨の範囲内において、種々の変形・変更が可能である。

【符号の説明】

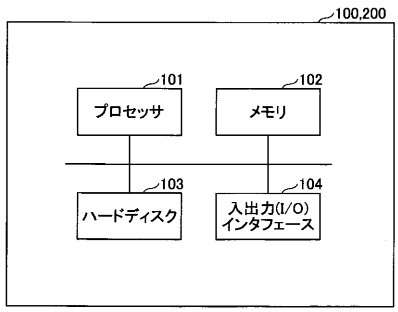
【００９０】

１０　ニューラルネットワーク構造
２０　テキストモダリティニューラルネットワーク
３０　音声モダリティニューラルネットワーク
４０　共通ニューラルネットワーク
１００　学習装置
２００　音声合成装置

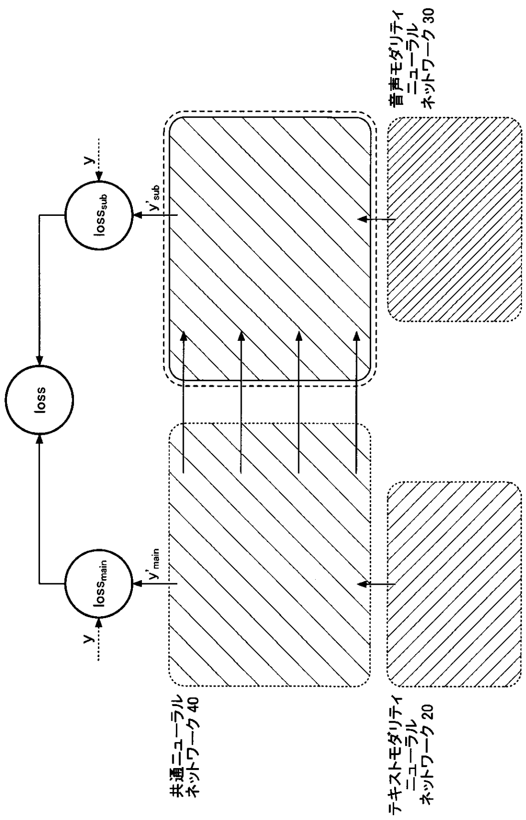
【図１】



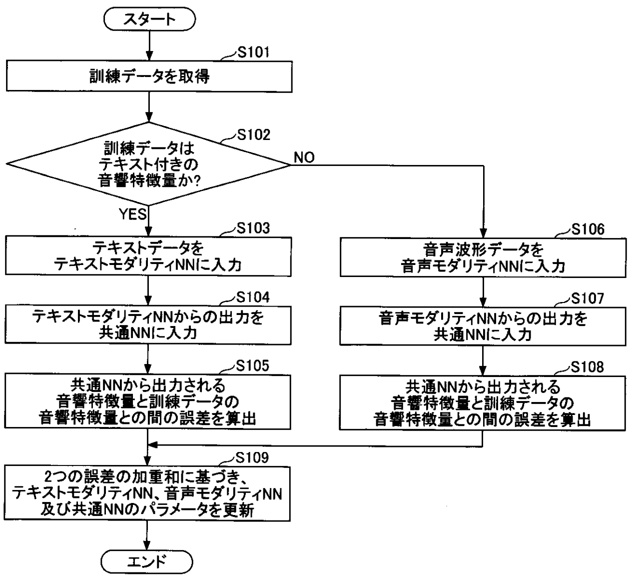
【図２】


100,200
101
プロセッサ
102
メモリ
103
ハードディスク
104
入出力(I/O)
インタフェース

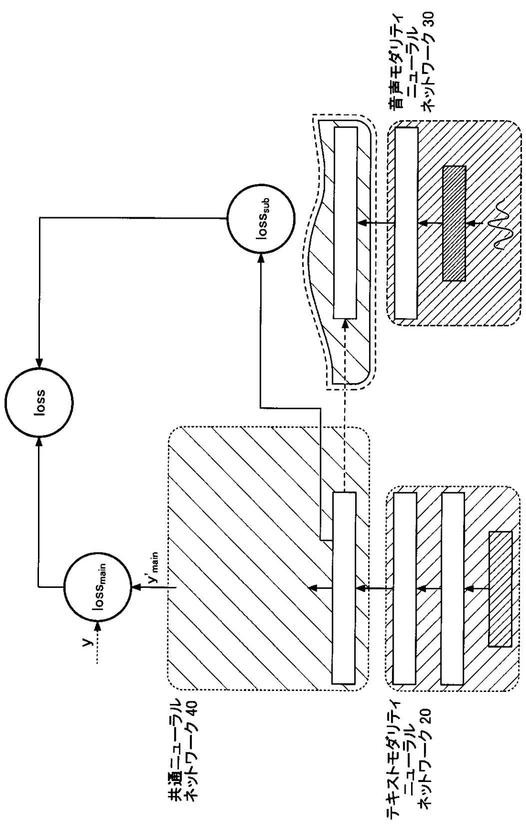
【図３】



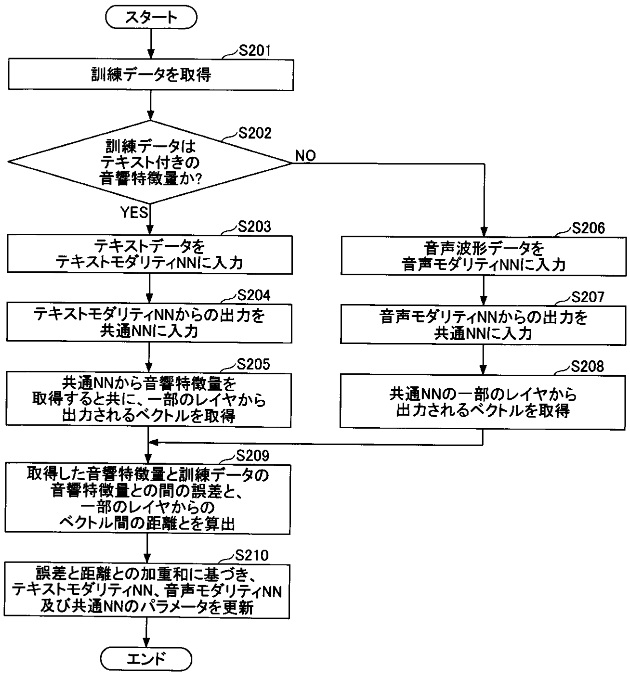
【図４】


スタート
S101
訓練データを取得
S102
訓練データは
テキスト付きの
音響特徴量か?
NO
YES
S103
テキストデータを
テキストモダリティNNに入力
S104
テキストモダリティNNからの出力を
共通NNに入力
S105
共通NNから出力される
音響特徴量と訓練データの
音響特徴量との間の誤差を算出
S106
音声波形データを
音声モダリティNNに入力
S107
音声モダリティNNからの出力を
共通NNに入力
S108
共通NNから出力される
音響特徴量と訓練データの
音響特徴量との間の誤差を算出
S109
2つの誤差の加重和に基づき、
テキストモダリティNN、音声モダリティNN
及び共通NNのパラメータを更新
エンド

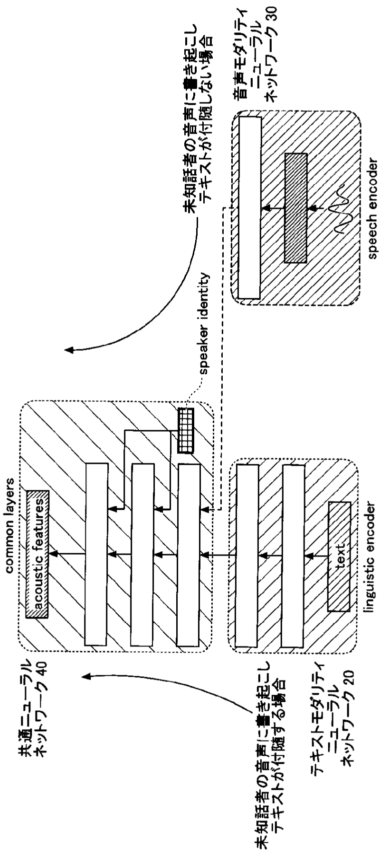
【図５】


音声モダリティ
ニューラル
ネットワーク30
loss_sub
loss
loss_main
y'_main
y
共通ニューラル
ネットワーク40
テキストモダリティ
ニューラル
ネットワーク20

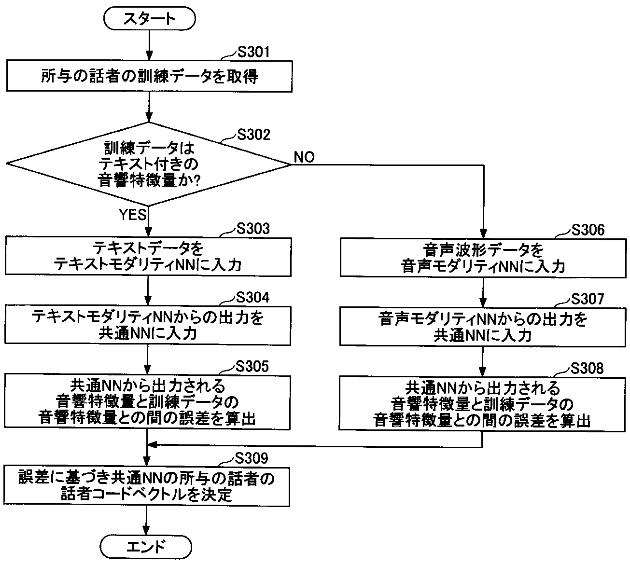
【図６】


スタート
S201
訓練データを取得
S202
訓練データは
テキスト付きの
音響特徴量か?
YES
NO
S203
テキストデータを
テキストモダリティNNに入力
S204
テキストモダリティNNからの出力を
共通NNに入力
S205
共通NNから音響特徴量を
取得すると共に、一部のレイヤから
出力されるベクトルを取得
S206
音声波形データを
音声モダリティNNに入力
S207
音声モダリティNNからの出力を
共通NNに入力
S208
共通NNの一部のレイヤから
出力されるベクトルを取得
S209
取得した音響特徴量と訓練データの
音響特徴量との間の誤差と、
一部のレイヤからの
ベクトル間の距離とを算出
S210
誤差と距離との加重和に基づき、
テキストモダリティNN、音声モダリティNN
及び共通NNのパラメータを更新
エンド

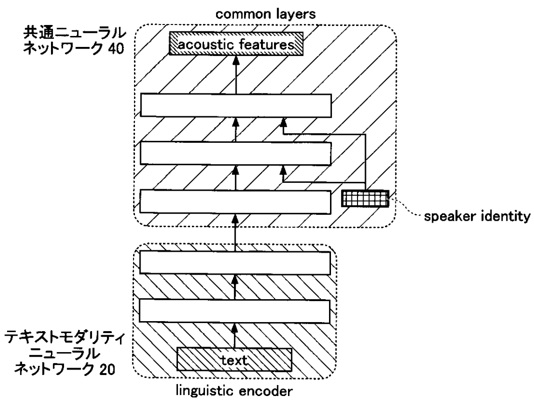
【図７】


未知話者の音声に書き起こし
テキストが付随しない場合
音声モダリティ
ニューラル
ネットワーク30
speech encoder
speaker identity
common layers
acoustic features
text
linguistic encoder
共通ニューラル
ネットワーク40
未知話者の音声に書き起こし
テキストが付随する場合
テキストモダリティ
ニューラル
ネットワーク20

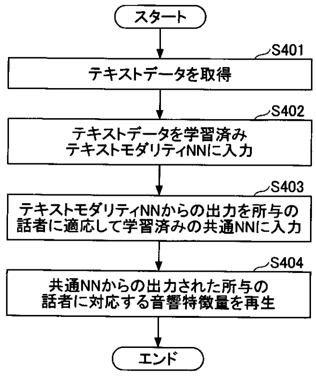
【図８】


スタート
S301
所与の話者の訓練データを取得
S302
訓練データは
テキスト付きの
音響特徴量か?
YES
NO
S303
テキストデータを
テキストモダリティNNに入力
S304
テキストモダリティNNからの出力を
共通NNに入力
S305
共通NNから出力される
音響特徴量と訓練データの
音響特徴量との間の誤差を算出
S306
音声波形データを
音声モダリティNNに入力
S307
音声モダリティNNからの出力を
共通NNに入力
S308
共通NNから出力される
音響特徴量と訓練データの
音響特徴量との間の誤差を算出
S309
誤差に基づき共通NNの所与の話者の
話者コードベクトルを決定
エンド

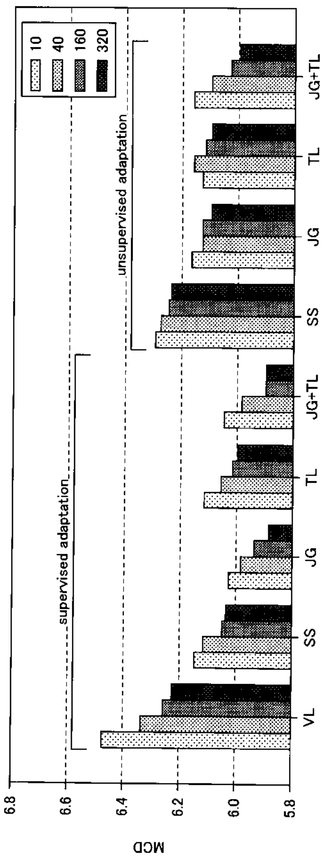
【図９】


common layers
共通ニューラル
ネットワーク 40
acoustic features
speaker identity
テキストモダリティ
ニューラル
ネットワーク 20
text
linguistic encoder

【図１０】


スタート
S401
テキストデータを取得
S402
テキストデータを学習済み
テキストモダリティNNに入力
S403
テキストモダリティNNからの出力を所与の
話者に適応して学習済みの共通NNに入力
S404
共通NNからの出力された所与の
話者に対応する音響特徴量を再生
エンド

【図１１】


10
40
160
320
supervised adaptation
unsupervised adaptation
VL
SS
JG
TL
JG+TL
SS
JG
TL
JG+TL
6.8
6.6
6.4
6.2
6.0
5.8
MCD